turn of the spring and summer sun effaces it. Furthermore, the rushing of the colder air to this region leaves the warmer air to occupy the other adjacent regions. In summer it would seem that the excessive heating of certain areas acts in a similar manner. It would also seem that cloudy weather, contrary to the general belief, perpetuates cold, at least in mountain regions. For instance, in clear years it is not cold in the Navajo country. This winter, a very cloudy one, the mercury has gone as low as 35 degrees below zero. Also in the cloudy, rainy Olympic country of Washington there are extensive glaciers, though the mountains, even in highest points, do not exceed 8,000 feet in elevation; while mountains in the drier regions of Washington, Idaho and Montana in the same latitude, though of practically the same elevation, possess no glacier fields. It appears that if there was no more precipitation in western Washington than there is in, say, Arizona, the glaciers of the Olympic mountains would not exist.

The writer hopes that others who have better facilities for observation will look further into this subject.

Scientific Measurement of the Achievements of Pupils.

F. J. KELLY, Dean School of Education, University of Kansas.

Advancement of any science depends primarily upon the accuracy with which the materials entering into that science can be measured. Until education can measure its products more accurately than it yet does, the claim that education is a science will not be generally allowed. It is the purpose of this paper to indicate some of the steps which have been taken to make possible a more accurate measurement of pupils' achievements.

Before telling of these recent efforts in deriving more accurate measures we must set forth quantitatively the extent of inaccuracy which prevails in our ordinary measures of educational products. We are constantly measuring the results of instruction. Examinations have been a part of school procedures ever since schools existed. On the basis of these examinations honors are awarded, pupils are encouraged to think that they are brilliant, or discouraged to think that they are stupid. Civil-service positions are awarded; teachers are granted certificates to teach; in fact, very much of our social structure rests upon examinations, which are the present-day measures of achievement. How reliable or unreliable these measures are is not generally known among people, even those engaged in educational work. There is just a sort of vague feeling that they are not a very satisfactory means of determining achievement.

The three most typical studies revealing the extent of the reliability of the examination paper as a measure are (1) that of F. Y. Edgeworth, professor of political economy in the University of Oxford; (2) those of Starch and Elliott, of the University of Wisconsin; and (3) that of the writer. Professor Edgeworth raised the question of the validity of the civil-service examinations in England, and in order to measure the reliability of the ratings upon the civil-service examination papers he sent facsimile reproductions of one of the examination papers to a group of twenty-eight head masters of the schools from whom examiners were chosen. Any one of the twenty-eight head masters was admitted by the civil-service commission to be competent to rate

the paper, and might have been chosen as the one to rate the papers used by the commission in selecting civil-service candidates. These twenty-eight examiners marked this paper from 45 to 100, seven of them marking it 72.5 or below and seven others marking it 85 or above. With this as a beginning, Professor Edgeworth made a very extended study of civil-service examinations, and gave as his conclusion the following: "I find the element of chance in these public examinations to be such that only a fraction, from one-third to two-thirds, of the successful candidates can be regarded as quite safe—above the danger of coming out unsuccessful if a different set of equally competent judges had happened to be appointed."

Starch and Elliott, in the University of Wisconsin, sent out facsimile reproductions of an English examination paper and of a geometry examination paper to the heads of English and mathematics departments, respectively, of the high schools of the North Central Association of Colleges and Secondary Schools. These schools represent the highest type of schools in the whole upper Mississippi valley and these department heads are persons who have been long in the service and have had a chance to discover the standards which prevail in examination papers in their departments. The English paper was rated by 142 English teachers and the geometry paper was rated by 116 mathematics teachers. Of these 142 English teachers 91 were in schools where 75 was the pass mark, and the median mark given by these teachers was 88.3, with a median deviation of 4.5 points. The other 51 teachers were in schools using 70 as a pass mark, and they gave a median mark of 87.2, with a median deviation of 4.2 points. The 116 mathematics teachers gave to the geometry paper a median mark of 70, with a median deviation of 7.5 points. Some of the teachers marked the English paper below 65, while a goodly proportion of them marked it above 90; 37 of the 116 mathematics teachers marked the geometry paper below 65, while 9 of them marked it above 85. This means that in the case of the English paper there is one chance in four that the mark would be changed by as much as 8.3 points when taken from the head of one English department to the head of another English department in these North Central Association schools, whereas there is one chance in four that the geometry paper would be changed by 15 points if it were taken from the head of one mathematics department to the head of another mathematics department for rating.

Such data as the foregoing are accumulating very rapidly. The writer undertook a few years ago to gather much more extensive data upon the subject of the extent of disagreement in values assigned to examination papers by two presumably competent judges. New York state has had a system of regents' examinations for more than thirty years. These examinations have been used during all this time as the basis for promotion in all the accredited high schools of the state. They were formerly given annually, and have for many years past been given semiannually under the direction of the teachers of the high schools, and the papers have been rated first by the teachers and then by the group of examiners employed by the regents. Because of the long standing of this system it appears that if persons can come to an agreement as to the value of an examination paper, such agreement should have been reached in New York state. In 1914 I studied the markings given, first by teachers and second by the regents' examiners, to all the regents' examination papers written in the high schools of New York in the year 1912.

This gave a total of 392,352 papers. Of these papers 18.5 per cent were marked failed by the teachers. When the papers which were marked passed by the teachers were sent in to the regents an additional 15.7 per cent of these were marked failed by the regents' examiners. It will appear from this, in the first place, that the disagreement between the teachers and the regents as to what constitutes a passing paper is very marked. In some of the subjects the percentage failed by the regents of those passed by teachers was much higher than 15.7 per cent. For example, of all the mathematics papers, 25.7 per cent of those passed by the teachers were marked failed by the regents. In commercial subjects 20.9 per cent of those passed by teachers were failed by the regents. The distribution of the teachers' ratings upon those papers which the regents subsequently marked failed is even more illuminating. Although the passed mark used by teachers and regents is 60 in New York state, 6.4 per cent of all the papers which the regents subsequently failed were marked 79 or better by the teachers.

Turning now to those papers which both the teachers and the regents allowed to pass, it is interesting to note the range of difference between the mark given a paper by the teacher and the mark given the same paper by the regents. Of all the papers marked 75 by the teachers only 7.48 per cent were given the same mark by the regents; 25 per cent of the papers marked 75 by the teacher were increased one point or more by the regents; 25 per cent were decreased seven points or less. It appears, therefore, that there is one chance in four that a paper marked 75 by the teacher will be marked as low as 68 or lower by the regents. When this is considered in connection with the rather narrow range in which marks lie (less than 25 per cent of the marks being above 74 on all the papers), the significance of this disagreement becomes apparent.

No more details of this study need be given in such a paper as this. If physicians had no better means to determine temperature and measure results of their prescriptions than we are using in education to measure the results of our practices, we would hesitate to call medicine a science. If pharmacists could not agree better on the measures of their drugs, if engineers could not agree better on the measures of their electric currents, we would cease to have much confidence in the science of pharmacy or in the science of engineering. It is because of its bearing upon this very fundamental problem that I am keenly interested in the development of more scientific measures of educational products.

In the more progressive educational circles more accurate measures in certain school subjects are now available. In the fundamentals in arithmetic, for example, the Courtis tests are used very widely to measure the degree of skill which children possess in the abilities of addition, subtraction, multiplication and division.

By means of these tests, given and scored according to directions, it is possible to determine the relative standing of individual children or groups of children in these fundamentals. As a matter of fact, these tests have been given sufficiently widely so that it is possible to compare the achievements of children in cities of the first class with children in cities of the second class and cities of the third class and in rural schools in these important fundamental operations. It is also possible to measure the increase in achievement

which children make in these fundamentals under the instruction of a given teacher, and thus measure the teacher's efficiency in this line with almost entire accuracy.

In the same fashion progressive schools now define achievement in penmanship according to the comparison of the child's writing with certain samples whose excellence has been determined by a large number of competent judges. Thus when a child is able to write so as to achieve a mark of 50 by the Kansas City scale and another child is able to write so as to achieve a mark of 60 by the same scale, we are reasonably certain that the second child writes about as much better than the first child as sample 60 is better than sample 50 on the scale. We can by means of the scale measure the improvement of children's writing under certain methods of teaching penmanship. We can measure the success of the teacher by determining how much progress her children make in penmanship under her instruction. We know from wide use of the scale how much progress children on the average make in each grade, and we can compare the achievement of a given teacher with average achievement of teachers in a like grade.

There are also tests which may be applied to measure efficiency in oral reading and efficiency in silent reading. In the former the score will indicate whether the reader is poor in pronunciation, poor in expression or poor in his knowledge of words. In silent reading separate measures are taken of the understanding of the printed paragraph and of the speed with which a child reads. All of these tests are so devised that there is little or no difference in rating by several judges of the same examination paper. When we find, as we have found, for example, that Iowa children, on the whole, read better than Kansas children to an average extent of 10 per cent of the number of words per minute which can be read with complete understanding, it is a significant discovery, and it has been used as a basis for the demand that Kansas children be allowed to have in their public-school work more than one reader per year. By being able to make all of these comparisons, from child to child or group to group, we are able to measure the efficiency of certain practices in elementary education.

I do not assume that you are interested in the details of these tests and scales now being devised. You are interested in the effort being made by men in education to derive such standard definitions of achievement as will enable us to build a science to replace the guesswork of former days. Suffice it to say that practically all leading cities in the country are now employing directors of educational tests, whose business it is to interpret the materials and methods employed by the teachers in terms of these more scientific definitions of achievement.

The tests which I have mentioned refer only to the tool subjects, supposed to be mastered in the elementary school. When it comes to information subjects, such as relate to vocational courses, or even to a good deal of the work in the academic departments of high school and college, the development has been even more slow. It is certainly true, however, that definitions of units in both high school and college must be clarified before we can hope for recognition.